The Effect of Selection on Genealogies

N. H. Barton*[,1] and A. M. Etheridge[†]

*Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom and
[†]Department of Statistics, University of Oxford, Oxford OX1 3T6, United Kingdom

Manuscript received October 30, 2002
Accepted for publication November 10, 2003

ABSTRACT

The coalescent process can describe the effects of selection at linked loci only if selection is so strong that genotype frequencies evolve deterministically. Here, we develop methods proposed by Kaplan, Darden, and Hudson to find the effects of weak selection. We show that the overall effect is given by an extension to Price's equation: the change in properties such as moments of coalescence times is equal to the covariance between those properties and the fitness of the sample of genes. The distribution of coalescence times differs substantially between allelic classes, even in the absence of selection. However, the average coalescence time between randomly chosen genes is insensitive to the current allele frequency and is affected significantly by purifying selection only if deleterious mutations are common and selection is strong (*i.e.*, the product of population size and selection coefficient, $Ns > 3$). Balancing selection increases mean coalescence times, but the effect becomes large only when mutation rates between allelic classes are low and when selection is extremely strong. Our analysis supports previous simulations that show that selection has surprisingly little effect on genealogies. Moreover, small fluctuations in allele frequency due to random drift can greatly reduce any such effects. This will make it difficult to detect the action of selection from neutral variation alone.

WE develop a diffusion approximation, first introduced by Kaplan *et al.* (1988), which extends the coalescent to take account of arbitrary forms of selection. Kingman (1982) introduced the coalescent process as a simple description of the genealogical relationships among a set of neutral genes. Although the theory of the coalescent has developed largely independently, it is closely related to the classical concept of identity by descent (Nagylaki 1989). The coalescent extends naturally to describe structured populations, in which genes may be found in different places or embedded in different genetic backgrounds. The effects of selection can easily be included, provided that it is so strong relative to random drift that the frequencies of different genetic backgrounds can be approximated as changing deterministically (*i.e.*, the product of population size and selection coefficient, $Ns \gg 1$; Kaplan *et al.* 1988; Hudson 1990).

In some cases, assuming that the genetic or spatial structure of a population changes deterministically is a good approximation. For example, when a single favorable mutation arises and spreads, it carries with it any linked variants. The effects of such "selective sweeps" on genetic variability can be accurately described by assuming that the new allele increases exponentially, even though it is subject to strong random fluctuations in the early generations, when it is present in few copies (Maynard Smith and Haigh 1974; Kaplan *et al.* 1988; Barton 1998). However, the deterministic approximation plainly fails when selection is weak or absent. Consider the relationships between genes that can be of two allelic types. Even if these alleles are neutral, and so represent an arbitrary labeling of the genes, two genes of the same allelic type are likely to be substantially more closely related than are two genes of different type. Moreover, the average relationship between randomly chosen genes depends on the allele frequency, since an allele that happens to have increased by chance will cause a selective sweep just as if it had increased by selection. Although relationships averaged over the distribution of allele frequencies and over allelic classes must be unaffected by the labeling of neutral alleles, relationships do depend on allelic class and on allele frequency (*e.g.*, Slatkin 1996).

We usually do not know which alleles are selected and so can observe relationships only among randomly chosen genes in populations with random genotype frequencies. However, with selection, even the average relationships are distorted and can be calculated using the structured coalescent only when selection is much stronger than drift. Weak selection ($Ns \sim 1$) spread across many loci can have significant cumulative effects (McVean and Charlesworth 2000). Even when Ns is large, fluctuations may still be important. For example, Barton and Navarro (2002) showed that the effects of balancing selection at multiple loci are strongly affected by drift even when Ns is extremely large, because each particular genetic background is present in few (if any) copies.

[1] *Corresponding author:* Institute of Cell, Animal and Population Biology, University of Edinburgh, W. Mains Rd., Edinburgh EH9 3JT, Scotland. E-mail: n.barton@ed.ac.uk

When selection and drift are of comparable strength, a purely coalescent-based approach becomes complicated. Neuhauser and Krone (1997) and Krone and Neuhauser (1997) have shown that certain kinds of selection can be represented by "ancestral graphs," which are constructed by allowing branching as well as coalescence as one moves back in time, followed by a culling of potential ancestors to generate the genealogy. This method is computationally demanding, especially with strong selection, because of the proliferation of ancestral lineages. Slade (2000a,b, 2001) and Fearnhead (2001) have introduced modifications that make calculations feasible for stronger selection. Nevertheless, this method does not seem likely to lead to a deeper analytical understanding, which would extend to more general kinds of selection and more complex genetics. Donnelly and Kurtz (1999a) and Slade (2001) have also shown how recombination can be included with selection in the algorithm. However, despite these various advances, the method is still computationally intensive for strong selection: for example, with overdominance, only $Ns < 10$ or so can be simulated (Slade 2000a). Moreover, it is limited to certain kinds of selection: linear frequency dependence or selection on diploids requires branching into three potential ancestors instead of two, and, more generally, a kth order polynomial dependence of haploid fitness on allele frequency requires branching into $(k + 2)$ virtual ancestors. In practice, anything beyond the simplest kind of epistasis or frequency dependence is ruled out.

Other recent simulation techniques follow the state of the whole population backward through time. Donnelly *et al.* (2001) discuss methods for importance sampling, which start with the well-understood neutral process, and apply a bias that represents the action of selection. This is most efficient when selection is weak. Slatkin (2001) begins by reversing the selective process, which should allow stronger selection to be represented accurately. Because the diffusion is reversible with additive selection, the procedure is exact in this case. With nonadditive selection, Slatkin (2001) uses a procedure that approximates the correct backward diffusion. One can also follow the evolution of population allele frequencies back through time and use the Metropolis-Hastings algorithm to sample the appropriate distribution of frequencies. Under the diffusion approximation, the probability of any particular path is given by a Gaussian distribution of velocities around their deterministic expectation, which approximates the product of Markov transition matrices (Schulman 1981; Rouhani and Barton 1987).

Kaplan *et al.* (1988) introduced a more direct approach to the problem, by following relationships between genes within and between allelic classes, conditional on the frequencies of those classes in the population. However, their approach has not been taken up. This may be partly because Kaplan *et al.* (1988) did not specify boundary conditions for their equations, so that they could not be solved using standard software; however, Darden *et al.* (1989) do provide an alternative numerical algorithm. Barton *et al.* (2003) give a rigorous justification for Kaplan *et al.*'s (1988) diffusion equations, including the necessary boundary conditions. In this article, we show that in the absence of selection, the overall average relationship between random genes in a random population is the same as under simple neutrality. This must of course be the case, since labeling a pair of neutral alleles cannot affect the distribution of genealogies. However, this result extends to give a general expression for the effect of selection on the genealogy, which can be seen as an extension of Price's (1970) equation.

The analysis is of a neutral locus that is linked to a single selected locus that carries two alternative alleles. (Setting the recombination rate to zero allows us to follow genealogies at a single selected locus). Essentially the same equations apply to probabilities of identity in state, assuming infinite-allele mutation at the neutral locus, the mean and higher moments of coalescence time, and the full distribution of coalescence times. The equivalence between these can be seen by noting that the identity in state is the generating function for the distribution of coalescence times. Our numerical results are mainly for the distribution of pairwise coalescence times, but in the last part of the article we consider the distribution of the total length of a large genealogy.

We begin by setting out the diffusion approximation for identities in allelic state; the equations for coalescence times are essentially the same. We then change variables to work with (i) the average over randomly chosen pairs of genes, (ii) differences associated with one or the other allele, and (iii) differences within classes *vs.* between classes. This change of variables leads to a simple formula for the expectation over the stationary distribution and to approximations for strong mixing between classes and for strong selection. Throughout this first part of the article, two simple examples are used to illustrate the derivations (two genes sampled from either a neutral locus or a locus under balancing selection). In the later sections, a wider range of parameters is explored.

THE MODEL

Consider selection on a single locus that carries two alleles, labeled P, Q. This is linked to a neutral second locus, with recombination rate r. Allele frequencies at the selected locus are p, q at the beginning of the generation. For simplicity, assume that selection acts on haploids; however, detailed assumptions about the life cycle do not affect the diffusion approximation. Numerical examples assume purifying selection with fitnesses of Q, P of $1{:}1 + s$; balancing selection is modeled by assuming frequency dependence such that s is replaced by $s(p_0 - p)$. This is close to a model of overdominance with

TABLE 1

Summary of notation

N	Effective population size	T	Scaled time, $t/2N$
μ	Sum of mutation rates, $\mu_{Q\to P} + \mu_{P\to Q}$	U	Scaled mutation rate, $N\mu$
$\overline{p}$	Equilibrium under mutation, $\mu_{Q\to P}/\mu$		
s	Selection favoring allele P	S	Scaled purifying selection, Ns
s_b	Strength of balancing selection	S_b	Scaled balancing selection, Ns_b
p_0	Equilibrium under balancing selection		
ν	Rate of mutation to new neutral alleles	V	Scaled neutral mutation rate, $N\nu$
r	Recombination rate	R	Scaled recombination rate, Nr
$f_{j,k}$	Identity in allelic state among j genes of type Q, k of type P	$\bar{f}$	Average identity, $\Sigma_{j=0}^{j+k} q^j p^k \binom{j+k}{j} f_{j,k}$
$J_{j,k}$	Expected total length of a genealogy relating j, k genes of type Q, P	$\bar{J}$	Expected total length, averaged over samples
		$\mathcal{T}$	Total length; expected value is J
$\Phi_{0,2}$, $\Phi_{1,1}$, $\Phi_{2,0}$	Distribution of pairwise coalescence times	τ	Mean pairwise coalescence time; $2\tau = \mathcal{T}$
$\tau_{2,0}$, $\tau_{1,1}$, $\tau_{0,2}$	Mean pairwise coalescence time	$\overline{\Phi}$	Distribution, averaged over samples
$E[\ ]$	Expectation over the stationary density	$\bar{\tau}$	Mean pairwise coalescence time; $2\bar{\tau} = J$
n	No. of genes in the sample	Δ_n	Half the regression of identity on allele frequency in the sample
		Δ_n^J	The same, but for total length
		Δ_n^τ	The same, but for mean pairwise coalescence time
$\mathcal{L}$	Differential operator, $2(U(p - \overline{p}) + Spq) + \frac{pq}{2}\partial_p$	θ	Difference in identity within *vs.* between classes, $f_{PP} - 2f_{PQ} + f_{QQ}$
		θ^τ	The same, for mean pairwise coalescence time

diploid fitnesses $1 - sp_0{:}1{:}1 - sq_0$, with $p_0 + q_0 = 1$. (The relation between models of overdominance and linear frequency dependence is discussed by Neuhauser 1999.) Mutation then occurs at a rate $\mu\overline{p}$ from Q alleles to P alleles and $\mu\overline{q}$ in the opposite direction. (In terms of the actual mutation rates $\mu = \mu_{Q\to P} + \mu_{P\to Q}$, $\overline{p} = \mu_{Q\to P}/\mu$; the equilibrium under mutation alone is $\overline{p}$.) Where we consider identity in allelic state, mutation to novel alleles occurs at a rate ν at the linked neutral locus. Diploid zygotes are formed by random union and undergo meiosis. Finally, $2N$ gametes are sampled to found the next generation. (We keep the convention that population size is $2N$ genomes, corresponding to N diploid individuals.) This discrete time model is defined in more detail by Barton *et al.* (2003), who also set out the analogous continuous-time Moran model. Our notation is summarized in Table 1.

In the limit where selection, drift, and mutation are weak, we can take a diffusion approximation to this model. We scale selection, mutation, and recombination relative to N, and time relative to $2N$, so that $T = t/2N$, $S = Ns$, $R = Nr$, $U = N\mu$, and $V = N\nu$. [Note that Barton *et al.* (2003) scale time relative to N generations.] Suppose that we sample n genes. We need to follow the probability $f_{j,k}$ that j genes of type Q, and k genes of type P, are identical in state at the neutral locus. We assume this neutral locus to be subject to infinite-alleles mutation at rate ν. $f_{j,k}$ is also the generating function for the distribution of total length of the genealogy and so can be used to find the distribution of the number of segregating sites.

Kaplan *et al.* (1988, Equation 20) provide a system of diffusion equations for $f_{j,k}$ without recombination. This system is extended to include recombination by Hudson and Kaplan (1988). These give

$$0 = -2V(j+k)f_{j,k} + \left(\frac{j(j-1)}{2}\frac{(f_{j-1,k} - f_{j,k})}{q} + \frac{k(k-1)}{2}\frac{(f_{j,k-1} - f_{j,k})}{p}\right) + 2\left(jp\left(U\frac{\overline{q}}{q} + R\right)(f_{j-1,k+1} - f_{j,k}) + kq\left(U\frac{\overline{p}}{p} + R\right)(f_{j+1,k-1} - f_{j,k})\right) + 2(U(\overline{p} - p) + Spq)\frac{\partial f_{j,k}}{\partial p} + \frac{pq}{2}\frac{\partial^2 f_{j,k}}{\partial p^2} \qquad (1)$$

with $f_{0,1} = f_{1,0} = 1$ by convention. Barton *et al.* (2003) give a rigorous derivation for this stationary version, for $n = 2$.

The terms involving $V = N\nu$ represent the steady decay in identity due to mutation at the neutral locus. The positive terms $(f_{j-1,k} - f_{j,k})/q$, $(f_{j,k-1} - f_{j,k})/p$ represent the increase in identity due to coalescence within allelic classes. The terms involving differences in identity $(f_{j-1,k+1} - f_{j,k})$, $(f_{j+1,k-1} - f_{j,k})$ represent the movement of genes between allelic classes by mutation of allelic classes and by recombination. Note that mutation from allele Q to allele P dominates recombination when allele

P is rare [term $U(\bar{p}/p) + R$ in Equation 1], because most copies of *P* will in that case have arisen as recent mutations from *Q*. Finally, the last two terms represent the proportion of populations currently at allele frequency p that derive from populations with a different frequency; this process is approximated as a backward diffusion.

In principle, the full distribution of genealogies can be recovered by assigning a notional mutation rate to each node, $\{\nu_i, \nu_{ij}, \nu_{ijk}, \ldots\}$ say, and by following identities among *sets* of genes that are either in background *Q* or *P* ($f_{\{a\},\{b,c\}}$, say). Then, terms such as $j(j-1)/2$ in Equation 1 separate out into distinct terms, corresponding to different permutations of loci over backgrounds. This gives a set of equations in a very large number of variables and, worse, an extremely large number of *f*'s: all possible partitions of the lineages must be tracked separately. So, numerical solution is difficult for even three genes. This approach might be useful, however, for deriving simpler equations that describe particular features of the distribution of genealogies.

As detailed in Barton *et al.* (2003), the probabilities of identity are the minimal positive solutions to the equilibrium version of Equations 1. This implicitly specifies the boundary conditions for the system, but to obtain numerical solutions, we require them explicitly. Consider first small p. The right-hand side of Equations 1 is dominated by terms in $1/p$. Solving for these terms leads to

$$f_{j,k} = \frac{(k-1)f_{j,k-1} + 4N\mu\bar{p}f_{j+1,k-1}}{(k-1) + 4N\mu\bar{p}} \quad (k > 0)$$

$$f_{j,0} = \frac{j(j-1)f_{j-1,0} + 4N\mu\bar{p}\partial_p f_{j,0}}{j(j-1) + 4Nj\nu}. \tag{2}$$

Similarly, as p tends to 1, the terms in $1/q$ dominate and we obtain analogous boundary conditions to Equations 2.

Numerical methods for solving Equations 1 and 2 are explained in the appendix. Figure 1 gives a check on these methods for two genes. Three methods for calculating identities within and between neutral allelic classes are compared. First, identities can be calculated conditional on allele frequencies, simulated over 50,000 generations, for a population of $2N = 100$ (Figure 1, dots). Allele frequencies were simulated over 50,000 generations, using the exact backward transition matrix calculated for the Wright-Fisher model. Second, the identities can be calculated by solving the discrete equivalent of Equation 1, which gives exact results for the Wright-Fisher model. This involves linear equations for a set of three vectors for $f_{0,2}$, $f_{1,1}$, $f_{2,0}$, each of length $2N + 1$. The results fit closely with those estimated by simulation and are indistinguishable in Figure 1. Finally, the diffusion approximation (Equation 1) was used. This fits closely over most of the range (compare solid curves with dots). However, it underestimates identities between genes in rare allelic classes (*e.g.*, $f_{0,2}$ for $p \to 0$; Figure 1, top left). The approximation is expected to fail for $p \sim 1/N$, since only a small number of copies are involved. A similar comparison for $2N = 1000$ shows that the discrepancy is then restricted to a narrower region, as expected.

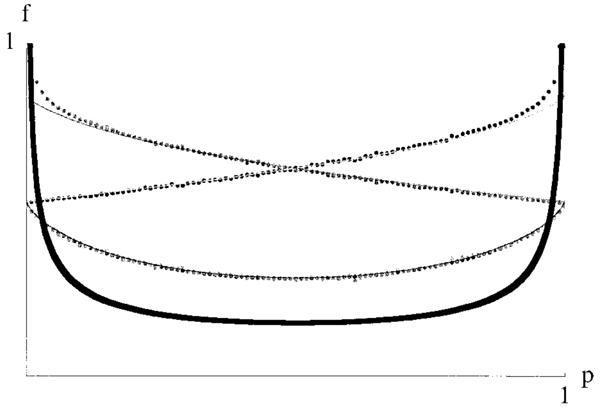

Figure 1.—Comparison among three methods for calculating identities as functions of allele frequency. The dots show identities calculated by simulation of a neutral allele over 50,000 generations and by solution of a matrix recursion; the values are barely distinguishable on this scale. The thin solid lines show the identities calculated using the diffusion approximation, Equations 1. The thick curve shows the stationary distribution. (This distribution is divided by 4 for clarity; the probability of being in the range $0.1 < p < 0.9$ is 0.59.) $2N = 100$, $\nu = \mu = 0.005$, $\bar{p} = 0.5$; thus, $U = V = 0.25$.

IDENTITIES, TREE LENGTHS, AND PAIRWISE COALESCENCE TIMES

First, consider probabilities of identity in allelic state under the infinite-alleles model. There are substantial differences between identities involving different allelic classes: identities between classes are much lower than those within (compare the lower curve for $f_{1,1}$ with the upper curves for $f_{0,2}, f_{2,0}$). Identities also vary substantially with allele frequency. Within-class identity decreases from ~1 when the allele is present in a few copies, down to $(1 + 4U\bar{p}f_{1,1})/(1 + 4U\bar{p})$ (Equation 2) when it is frequent enough for the diffusion approximation to hold ($p \gg 1/N$), and then down to a value somewhat greater than the neutral expectation of $1/(1 + 4V)$ when the allele nears fixation. The between-class identity necessarily approaches that within the commonest class near fixation and decreases in between, because there is then a rapid influx into the rarer class by mutation from the commoner class.

Note that there is considerable variation in the identities, and hence in the distribution of coalescence times, in any particular generation of a simulated population (Figure 2, top). This arises from the random history of allele frequencies (Figure 2, middle) and is an extra source of variation, over and above the intrinsic variation in the actual coalescence time. The latter is a sample

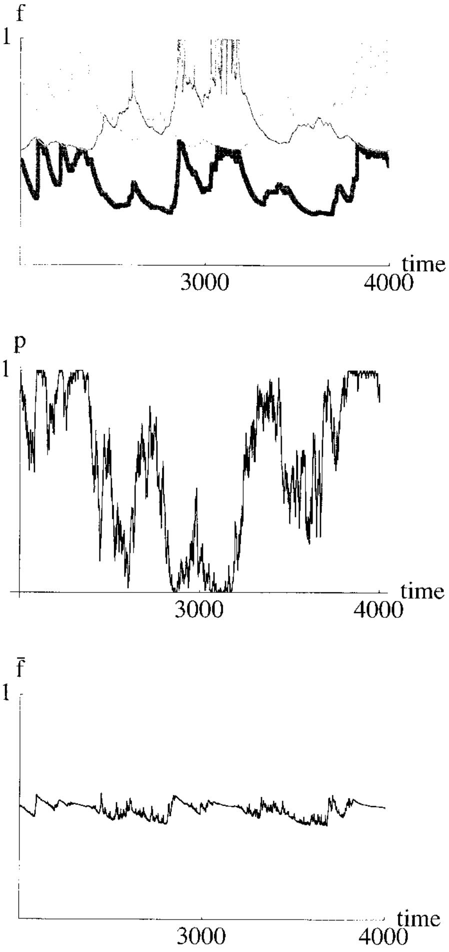

FIGURE 2.—An example of the time-course of identities (top), allele frequencies (middle), and average identity (bottom), from simulations of a neutral allele ($2N = 100$, $\nu = \mu = 0.005$, $\bar{p} = 0.5$, as in Figure 1). In the top, the bottom thick line shows the identity $f_{1,1}$ between allelic classes, over generations 2000–4000. The top two lines show the within-class identities $f_{2,0}$, $f_{0,2}$. Around generation 3000, allele P is lost (middle); then the identity $f_{0,2}$ is set to 1, and the identities $f_{1,1}$, $f_{2,0}$ become equal. The converse pattern is seen when allele Q is lost, around generations 2000 and 4000. Although the identities within and between classes fluctuate greatly with allele frequency, the average identity stays close to the expected value $\bar{f} = 1/(1 + 4N\nu) = 0.5$.

from a distribution that itself fluctuates with allele frequencies. The smooth curves shown in Figure 1 are the identities conditional on current allele frequency and are an average over the distribution of past fluctuations in allele frequency. Calculation of the variance in pairwise identity as a function of current allele frequencies would require consideration of associations among sets of four genes.

Figure 3 shows the same comparison as in Figure 1, but with balancing selection of strength $S_b = Ns_b = 4$ toward an equilibrium point $p_0 = 0.7$. Again, the three methods agree to high accuracy, except for identities within rare allelic classes. Even though selection is much stronger than mutation and drift, it has surprisingly little effect in reducing the identities $f_{2,0}$, $f_{1,1}$, $f_{0,2}$. The stationary distribution is now concentrated around $p_0 = 0.7$; because identities depend strongly on allele frequency, this might be expected to alter the identity between random pairs of genes, averaged over the stationary distribution. However, balancing selection reduces this average to only $E[\bar{f}] = 0.4977$, relative to the neutral value of $1/(1 + 4V) = 0.5$. This is because the average identity $\bar{f} = q^2 f_{2,0} + 2pqf_{1,1} + p^2 f_{0,2}$ is almost independent of allele frequency (*e.g.*, Figure 2, bottom). We consider this issue in more detail below.

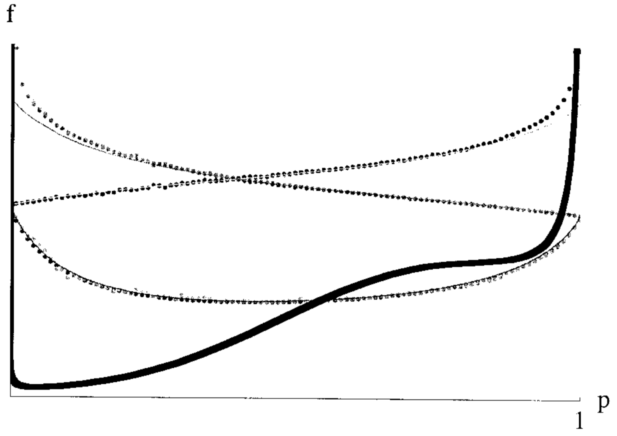

FIGURE 3.—Comparison among three methods for calculating identities as functions of allele frequency, under balancing selection. Parameters are as in Figure 1, except that there is balancing selection with coefficient $s(p_0 - p)$, $s = 0.08$, $p_0 = 0.7$. The stationary distribution is now concentrated around p_0; the probability of $0.1 < p < 0.9$ is increased to 0.74.

Similar equations can be derived for the probability density of total length of the genealogy, $\Phi_{j,k}$. This is a function of the total length, $\mathcal{T}$, and current allele frequency, p. After a long time, the density approaches a steady state, which satisfies

$$-(j + k)\frac{\partial \Phi_{j,k}}{\partial \mathcal{T}} = \left(\frac{j(j-1)}{2}\frac{(\Phi_{j-1,k} - \Phi_{j,k})}{q} + \frac{k(k-1)}{2}\frac{(\Phi_{j,k-1} - \Phi_{j,k})}{p}\right)$$
$$+ 2\left(jp\left(U\frac{\bar{q}}{q} + R\right)(\Phi_{j-1,k+1} - \Phi_{j,k}) + kq\left(U\frac{\bar{p}}{p} + R\right)(\Phi_{j+1,k-1} - \Phi_{j,k})\right)$$
$$+ 2(U(\bar{p} - p) + Spq)\frac{\partial \Phi_{j,k}}{\partial p} + \frac{pq}{2}\frac{\partial^2 \Phi_{j,k}}{\partial p^2}. \qquad (3)$$

At $\mathcal{T} = 0$, we have

$$\Phi_{j,0} = \frac{j(j-1)}{2q}$$
$$\Phi_{j,k} = 0 \qquad \text{for } j \neq k$$
$$\Phi_{0,k} = \frac{k(k-1)}{2p}. \qquad (4)$$

The boundary conditions at $\mathcal{T} = 0$ set the rate of coalescence within allelic classes as being inversely proportional to the frequency of the class. (Recall that time has been scaled relative to $2N$). The partial differential equations themselves have essentially the same form as

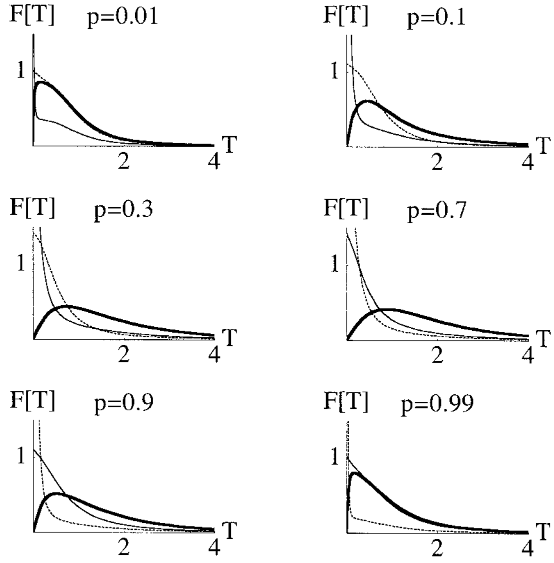

FIGURE 4.—The distribution of coalescence times, calculated from Equations 3, for a locus under balancing selection. The three parts show distributions at p = 0.01, 0.1, 0.5 (left to right). In each, the thick line shows the between-class distribution $\Phi_{1,1}$, the dashed curve shows $\Phi_{2,0}$, and the thin solid curve shows $\Phi_{0,2}$ ($U = 0.25$, $\overline{p} = 0.5$, $S = 4$, $p_0 = 0.7$, as in Figure 3). Time is scaled relative to $2N$ generations.

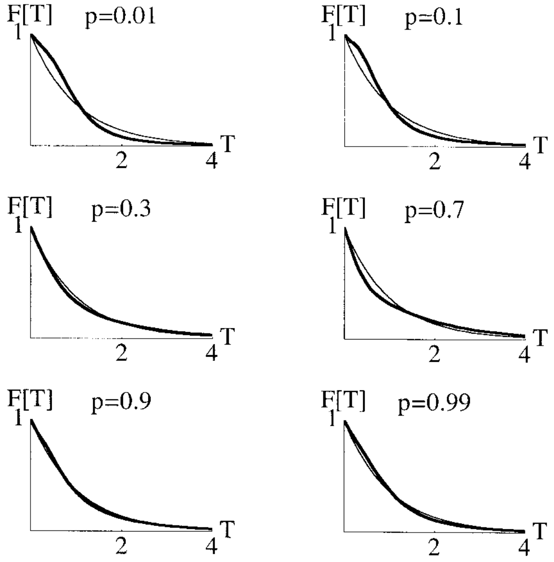

FIGURE 5.—The distribution of coalescence times between randomly chosen pairs of genes (thick line) compared with the neutral expectation Exp[$-T$] (thin line). Parameters are for a locus under balancing selection, as for Figures 3 and 4.

for the identities and represent the movement of genes between allelic classes and the diffusion of populations between allele frequency states.

Figure 4 shows the solution to Equations 3, for the same parameters as Figure 3. The rate of coalescence between allelic classes, $\Phi_{1,1}$, is necessarily zero, and so the distribution $\Phi_{1,1}$ passes through the origin (Figure 4, thick lines). However, when one or the other allele is rare, genes in the rarer class are likely to be descended from genes in the common class relatively recently. Hence, $\Phi_{1,1}$ rapidly approaches $\Phi_{2,0}$ when P is rare (Figure 4, top left and bottom right). Coalescence times within a rare allelic class are likely to be very short, unless the two genes derive from the common class via recent mutation. Because the mutational flux from common to rare is high, these two possibilities have comparable probability (see two terms $\sim 1/p$ in Equation 1). Thus, for small p the distribution $\Phi_{0,2}$ is a mixture of a singularity at zero and a component proportional to $\Phi_{2,0}$ (Figure 4, bottom curve in top left and bottom right). At intermediate frequency, both within-class distributions are close to the neutral expectation, Exp[$-T$] (Figure 4, middle).

Figure 5 compares the distribution of coalescence times between randomly sampled pairs of genes, $\overline{\Phi} = q^2\Phi_{2,0} + 2pq\Phi_{1,1} + p^2\Phi_{0,2}$, with the neutral formula Exp[$-T$], for the same example of balancing selection. When allele P is rare, the coalescence time tends to be shorter, while as the frequency approaches the intermediate value where the population is most likely to be found ($p \sim p_0 = 0.7$), coalescence times become slightly longer than the neutral expectation. As P approaches fixation, coalescence times again become slightly shorter than in the absence of selection and allelic structure. Overall, there is little change: in this example, balancing selection increases mean coalescence time by 13.9%.

The relation between Equations 1 and 3 can be understood by noting that the identity in state is the generating function for the distribution of coalescence times, with scaled parameter $4V$ (*i.e.*, $f = \int_0^\infty \exp[-4V\mathcal{T}]\Phi \mathbb{d}\mathcal{T}$. This can be confirmed by integrating the product of Equation 3 with Exp[$-4V\mathcal{T}$] over $\mathcal{T}$. The moments of coalescence times can be recovered by taking differentials of f at $V = 0$. For example, the expected total length $J = \int_0^\infty \mathcal{T}\Phi[\mathcal{T}]\mathbb{d}\mathcal{T}$ is given by

$$
\begin{aligned}
0 = n &+ \left(\frac{j(j-1)}{2}\frac{(J_{j-1,k} - J_{j,k})}{q} + \frac{k(k-1)}{2}\frac{(J_{j,k-1} - J_{j,k})}{p}\right) \\
&+ 2\left(jp(U\frac{\overline{q}}{q} + R)(J_{j-1,k+1} - J_{j,k}) + kq(U\frac{\overline{p}}{p} + R)(J_{j+1,k-1} - J_{j,k})\right) \\
&+ 2(U(\overline{p} - p) + Spq)\frac{\partial J_{j,k}}{\partial p} + \frac{pq}{2}\frac{\partial^2 J_{j,k}}{\partial p^2}
\end{aligned} \tag{5}
$$

with $J_{1,0} = J_{0,1} = 0$ by convention. As $p \rightarrow 0$, mean total length tends to

$$
J_{j,k} = \frac{(k-1)J_{j,k-1} + 4U\overline{p}J_{j+1,k-1}}{(k-1) + 4U\overline{p}} \quad (k > 0)
$$

$$
J_{j,0} = \frac{2}{j-1} + J_{j-1,0} + \frac{4U\overline{p}\partial_p J_{j,0}}{j(j-1)}. \tag{6}
$$

These limits can be found directly or from Equations 2. Note that for $n = 2$ genes, the expected total length

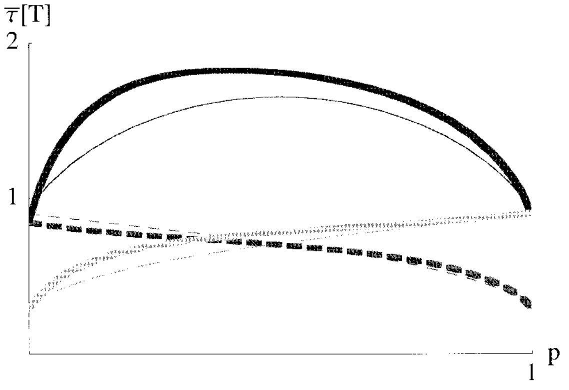

FIGURE 6.—Mean coalescence time within and between allelic classes, plotted against allele frequency. The thin lines are for neutral alleles, while the thick lines are for balancing selection $S = 4$, $p_0 = 0.7$; $U = 0.25$ as before. The top pair of curves are for genes in different allelic classes ($\tau_{1,1}$). The bottom two pairs are for mean coalescence time within classes (dashed curves, $\tau_{2,0}$; solid curves, $\tau_{0,2}$). Mean coalescence time between two P genes, $\tau_{0,2}$, decreases to $4U\bar{p}\tau_{1,1}/(1 + 4U\bar{p})$ as $p \rightarrow 0$; conversely, $\tau_{2,0} \rightarrow 4U\bar{q}\tau_{1,1}/(1 + 4U\bar{q})$ as $p \rightarrow 1$.

of the genealogy is twice the mean pairwise coalescence time, which we denote by τ.

Figure 6 shows how the mean coalescence time changes with allele frequency, with and without balancing selection. The mean coalescence time between genes in different allelic classes is much greater than that within classes, but approaches the same value as within the commoner class as that class approaches fixation. Within-class coalescence times approach $4U\bar{p}J_{1,1}/(1 + 4U\bar{p})$ as the allele becomes rare. This value is determined by a balance between the rapid rate of coalescence within rare classes and the rapid influx of copies by mutation from the commoner class.

A CHANGE OF VARIABLES

To make some approximations, and to understand average identity, it is helpful to change variables, as follows. First, consider the pairwise case. Let

$$\begin{aligned} \bar{f} &= q^2 f_{2,0} + 2pqf_{1,1} + p^2 f_{0,2}, & f_{2,0} &= \bar{f} - 2\,p\Delta + p^2\theta \\ \Delta &= -qf_{2,0} + (q - p)f_{1,1} + pf_{0,2}, & f_{1,1} &= \bar{f} + (q - p)\Delta - pq\theta \\ \theta &= f_{2,0} - 2f_{1,1} + f_{0,2}, & f_{0,2} &= \bar{f} + 2q\Delta + q^2\theta. \end{aligned} \qquad (7)$$

Similar transformations apply for the expected total length, J, and for the distribution of coalescence times, Φ. The average identity among randomly chosen pairs of genes is $\bar{f}$; Δ is the difference in identity between a gene associated with P and a random partner and a gene associated with Q and a random partner; and θ is the sum of the differences between identities of alleles in the same classes and alleles in distinct classes. Equations 1 transform to

$$0 = 1 - (1 + 4V)\bar{f} + \mathcal{L}(\partial_p\bar{f} - 4\Delta) + 4pqS\Delta \qquad (8a)$$

$$0 = -2\left(1 + 2V + U + U\frac{(q - p)}{pq}(\bar{p} - p) + R\right)\Delta + \mathcal{L}(\partial_p\Delta - 2\theta) + (p - q + 2Spq)\theta \qquad (8b)$$

$$0 = \frac{(1 - \bar{f})}{pq} - (3 + 4V + 4U + 4R + 4S(p - q))\theta + \mathcal{L}\left(\partial_p\theta + 2\frac{(p - q)}{pq}\theta\right) + \frac{(p - q)}{pq}(2\Delta - pq\partial_p\theta),$$

$$\text{where } \mathcal{L} = \left(2Spq + 2U(\bar{p} - p) + \frac{pq}{2}\partial_p\right). \qquad (8c)$$

Each variable is augmented by a diffusion term, $\mathcal{L}(\ )$. As is shown in Equation 9, the average of this diffusion term over the stationary distribution is zero; thus, it shifts the variable without producing a net change. The first equation shows that average identity is augmented by $4pqS\Delta[p]$, which is the product of the change in allele frequency due to selection, and the difference in identity between genes associated with P rather than with Q. Net identity increases if a higher identity is associated with an allele favored by selection. The crucial quantity, then, is Δ. The second equation shows that this is reduced by recombination and mutation, especially near the edges, and augmented by a term proportional to θ, which is the difference in identity within and between allelic classes. The last equation, for θ, shows that it is augmented by coalescence, especially at the boundaries, in proportion to $(1 - \bar{f})/pq$. It is also augmented by a term proportional to Δ.

The equations for the distribution of coalescence times are obtained by setting $V = 0$ and dropping the terms that do not involve $\bar{f}$, Δ, or θ (*i.e.*, 1 in Equation 8a and $1/pq$ in Equation 8c). Boundary conditions at $\tau = 0$ are that $\overline{\Phi} = 1$, $\Delta = 0$, and $\theta = 1/pq$. The expected total length (which is twice the mean pairwise coalescence time for $n = 2$) is given by the same equations as above, but setting $V = 0$, and subtracting $1/pq$ from Equation 8c.

Figure 7 shows the transformation for mean coalescence time, to the variables $\bar{\tau}$, Δ^τ, θ^τ (Table 1). The mean coalescence time between randomly chosen pairs of genes varies rather little with allele frequency and increases only moderately with increasing balancing selection (Figure 7, top). This is because the equation for $\bar{\tau}$ is driven mainly by the term $1 - \bar{\tau}$, which leads to the standard solution $\bar{\tau} = 1$. The diffusion terms $\mathcal{L}(\partial_p\bar{\tau} - 4\Delta^\tau)$ redistribute $\bar{\tau}$ across the range of allele frequencies, but, as we show below, do not alter its average value. The main driving term is small, because under balancing selection both S and Δ are zero somewhere in the interior, and so their product is small when allele frequency is concentrated in the interior.

Figure 8 shows the same comparisons, but for puri-

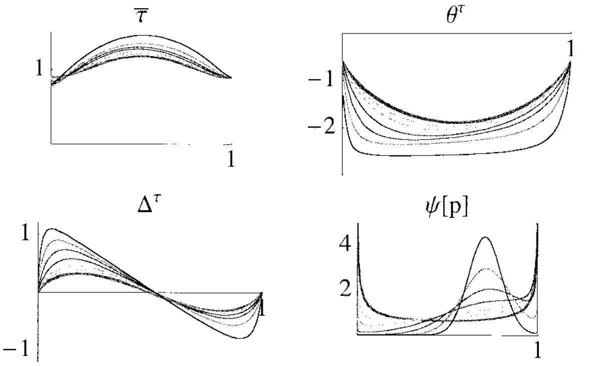

FIGURE 7.—Transformed representation of the mean coalescence time, for increasing strengths of balancing selection ($U = 0.25$, $p_0 = 0.7$). Top left, $\bar{\tau}$, the average over randomly chosen pairs of genes; bottom left, Δ^{τ}, the effect on mean coalescence time of association with P rather than with Q; top right, θ^{τ}, twice the difference in mean coalescence time within relative to between classes; bottom right, $\psi[p]$, the stationary distribution. The thick curves are for neutral alleles, and the successive thin curves are for $S = 1, 2, 4, 8, 16, 32$.

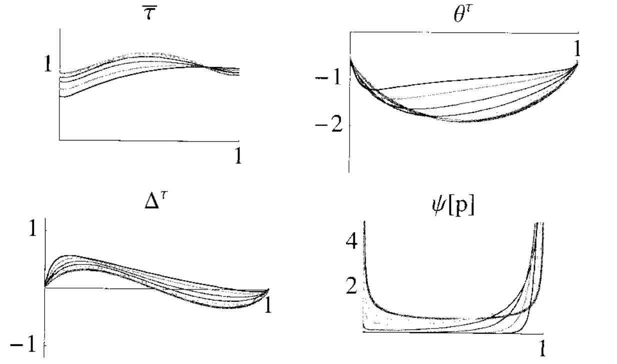

FIGURE 8.—Transformed representation of the mean coalescence time, for increasing strengths of purifying selection; parameters and notation are as in Figure 7. The thick curves are for neutral alleles, and the successive thin curves are for $S = 0.25, 0.5, 1, 2, 4, 8$.

fying selection. The overall mean $\bar{\tau}$ is affected little by weak selection ($S \leq 1$, say). With stronger purifying selection, mean coalescence time is substantially reduced when the favorable allele becomes rare (Figure 8, left side of top left), but is slightly increased when the favorable allele is common. The average identity is insensitive to selection when selection is weak because then Δ^{τ} changes sign in the center and so its product with Spq also changes sign; the net effect through the driving term $4Spq\Delta^{\tau}$ is thus small. As selection becomes stronger, Δ^{τ} becomes positive over a wider region, implying that the mean coalescence times involving genes in the favored allelic class become longer. However, the net effect on the pattern of $\bar{\tau}$ is hard to predict, because the diffusion term is strong. The next section sets out a simple result for the mean coalescence time, averaged over the stationary distribution, which necessarily does not include this diffusion term.

THE NET EFFECT OF SELECTION

A remarkably simple result is obtained by taking the expectation of the average identity over the stationary distribution, $E[\bar{f}]$. We know that the stationary distribution satisfies the forward diffusion $0 = (2Spq + 2(\bar{p} - p)U)\psi - \partial_p((pq/2)\psi)$. Integrating the first of Equations 8 over the stationary distribution ψ, we have

$$0 = 1 - (1 + 4V)E[\bar{f}] + \int \psi \mathcal{L}(\partial_p \bar{f} - 4\Delta)\, \mathrm{d}p + 4E[Spq\Delta[p]]. \tag{9}$$

Integrating by parts, the third term vanishes, and we have

$$E[\bar{f}] = \frac{1 + 4E[Spq\Delta[p]]}{1 + 4V}. \tag{10}$$

We immediately see that, in the neutral case ($S = 0$), $E[\bar{f}]$ is unaffected by the arbitrary labeling: it is given by the standard formula from the neutral theory, $1/(1 + 4V)$. Selection will perturb average identity by a proportion that depends on $E[Spq\Delta[p]]$; this will be small for purifying selection, since Δ changes sign somewhere in the center. However, it will be large and negative for balancing selection, if the null point of selection coincides with the null point for Δ. Then, both Δ and S will change sign near the center and so $E[Spq\Delta[p]]$ is negative throughout. Thus, balancing selection is expected to reduce average identity.

The mean coalescence time is given by the same equation as Equation 9, but with V set to zero, and $\Delta^{\tau} = p(\bar{\tau}_{0,2} - \bar{\tau}_{1,1}) + q(\bar{\tau}_{1,1} - \bar{\tau}_{2,0})$:

$$E[\bar{\tau}] = 1 + 4E[Spq\Delta^{\tau}[p]]. \tag{11}$$

Similarly, the distribution of coalescence times, averaged over random pairs of genes and over the distribution of allele frequencies, is

$$E[\bar{\Phi}[\tau]] = \mathrm{Exp}[-\tau] + 4\int_0^{\tau} \mathrm{Exp}[-(\tau - \tau')]E[Spq\Delta^{\Phi}[\tau', p]]\, \mathrm{d}\tau'. \tag{12}$$

The expected mean coalescence time can be calculated either directly or by using the right side of Equation 11. In numerical calculations, the latter is more accurate, both because it gives what is usually a small deviation from the neutral expectation of 1 and because the regions near fixation (where the stationary distribution diverges for $U\bar{p}$, $U\bar{q} < 1/4$) do not contribute significantly to the integral (since pq and Δ^{τ} tend to zero at the boundaries). The boundaries do not contribute to the deviation from neutrality because in these regions selection is negligible relative to drift, and there is rapid flow between backgrounds.

The relationships of Equations 9–12 extend to arbitrary numbers of genes in the sample. To be definite, consider the expected total length of the genealogy, $J_{j,k}$ (Equation 5). However, the same argument applies to other properties of the genealogy, such as the identity in state or the distribution of times to the most recent common ancestor. As mentioned above, one can write down the generating function for the complete density in the same form, and so this result applies for any quantities derived from it.

Summing over the binomial probability density for the number of genes $\{j, k\}$ sampled from each background, we obtain

$$0 = n + \frac{n(n-1)}{2}(\bar{J}_{n-1} - \bar{J}_n) + 4Spq\Delta_n^J + \mathcal{L}(\partial_p\bar{J}_n - 4\Delta_n),$$
$$\text{where } \bar{J}_n = \sum_{j+k=n} q^j p^k \binom{n}{j} J_{j,k} \quad \text{and} \quad \Delta_n^J = \frac{1}{2pq}\sum_{j+k=n} q^j p^k \binom{n}{j} (k - np) J_{j,k}. \tag{13}$$

The first term n represents the increase in tree length at a rate n, when n lineages are present. The second term represents coalescence at a rate $n(n-1)/2$; when coalescence occurs, the tree length expected among $(n-1)$ genes replaces that among n genes: $(\bar{J}_{n-1} - \bar{J}_n)$. The third term represents the perturbation to average tree length caused by selection. It is proportional to Δ_n^J, which is half the regression of average tree length on the number of P alleles in the sample (*i.e.*, of $J_{j,k}$ on k). The last term represents the effect of changes in allele frequency through time.

Taking the expectation over the stationary distribution, the last term vanishes, and we obtain a recursion:

$$E[\bar{J}_n] = E[\bar{J}_{n-1}] + \frac{2}{n-1} + \frac{2}{n(n-1)}E[4Spq\Delta_n^J]$$
$$= 2\sum_{j=1}^{n-1}\frac{1}{j} + E\left[4Spq\sum_{j=2}^{n}\frac{2}{j(j-1)}\Delta_j^J\right]. \tag{14}$$

We see immediately that in the absence of selection the expected total tree length is given by the neutral formula, $E[\bar{J}_n] = 2\Sigma_{j=2}^{n-1}(1/j)$. The perturbation due to selection during the time when there are j lineages in the genealogy is proportional to $2/j(j-1)$, which is the expected time for which j lineages persist before coalescence. This suggests that the effects of selection will primarily be on the deeper parts of the genealogy (*i.e.*, small j); however, it is not clear how Δ_j changes with j. Numerical calculations confirm our intuition that the effects of selection decrease rapidly as j increases. Figure 9 shows the successive perturbations due to selection for up to $n = 50$ genes, compared with the successive contributions $2/j$ under neutrality. The Δ_j decrease for $j > 10$, and so the net perturbation, which is further reduced by the factor $2/j(j-1)$, quickly becomes negligible. The sum of the perturbations is -0.29, compared with a neutral expectation of $E[\bar{J}] = 8.96$. Thus, purifying selection reduces total tree length by only 3.3% in this example, and more than half of that effect arises during the time when only two or three lineages are extant. (As explained in the APPENDIX, we use the insensitivity of the genealogy to allele frequency fluctuations in our calculations of terms involving large numbers of genes.)

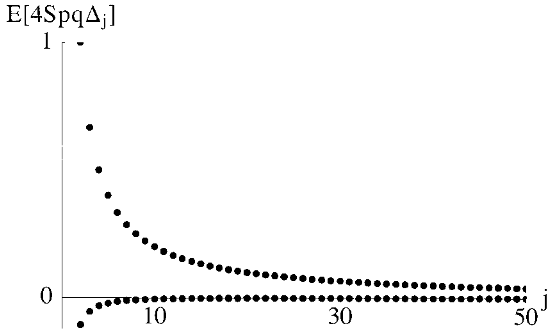

FIGURE 9.—The perturbations to total expected tree length caused by purifying selection, plotted against the number of extant genealogies, j (bottom series of points). These are calculated from $E[4pqS\Delta_j]$ (Equation 16). The top series of points show the contribution to total expected tree length expected under neutrality, $2/j$. Mutation rate is $U = 0.5$, $\bar{p} = 0.5$, and purifying selection $S = 2$; there is no recombination.

Equation 14 can be interpreted as an instance of ROBERTSON's (1966) "secondary theorem of natural selection," which states that the increase in any quantity caused by selection is equal to the covariance between that quantity and fitness. [This was independently developed into a general representation of selection by PRICE (1970).] In terms of the original variables, $E[4Spq\Delta_n^J] = E[\Sigma_{j+k=n}q^jp^k\binom{n}{j}S(k-np)J_{j,k}]$, which is equal to the covariance between the fitness of the sample of n genes and the expected total length of the genealogy that relates these genes. (Here, relative fitness of a sample with k copies of allele P is sk, or $2Nsk = 2Sk$ when rescaled.) Equation 14 shows that this covariance is exactly equal to the increase in the expected total length caused by selection. Because essentially the same equations apply to the identity f, which can be viewed as the generating function of the genealogy, we can see that relations similar to Equation 14 apply to any quantities that are linear functions of the probability density of genealogies (for example, the kth moments of coalescence times).

APPROXIMATIONS

There does not appear to be an explicit solution to Equations 1 or 8, even in the absence of selection. We therefore explore several approximations, in the hope that these might extend to more complex situations, which are difficult to investigate numerically. We consider in turn the extremes of no mixing between allelic classes, assuming that change in allele frequency is negligible and assuming that there is rapid mixing between backgrounds by mutation and recombination.

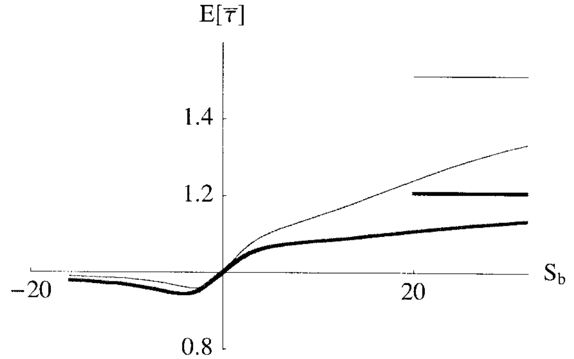

FIGURE 10.—The effect of balancing selection on the mean coalescence time, $E[\bar{\tau}]$. The strength of balancing selection, S_b, increases to the right; negative values correspond to disruptive selection; $p_0 = 0.7$. The two curves are for mutation rate $U = 0.5$ (thick curve) and 0.25 (thin curve). There is no recombination ($R = 0$). The straight lines are the predictions assuming that allele frequency is fixed at p_0.

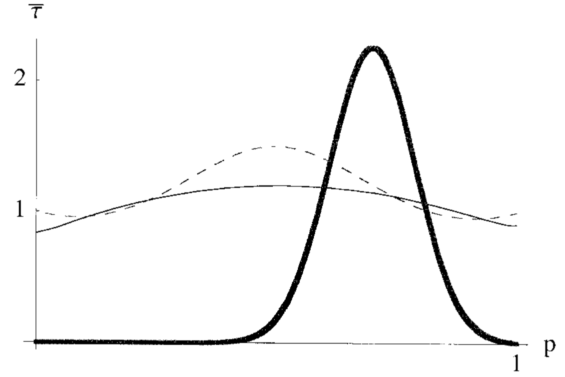

FIGURE 11.—The mean coalescence time, $\bar{\tau}$, as a function of allele frequency; $U = 0.5$, $\bar{p} = 0.5$, $R = 0$, $S_b = 32$, $p_0 = 0.7$. The thin solid curve gives the exact solution from Equations 5. This is compared with the prediction from Equation 15 (dashed line), which assumes that allele frequency is fixed; this prediction is independent of selection. The thick curve shows the stationary distribution.

No movement between backgrounds: If there is no recombination or mutation between alleles (U, $R = 0$) then the three identities given by Equations 1 change independently. The between-class identity $f_{1,1}$ clearly tends to zero, while the within-class identities are those of two separate populations of size p, q, which fluctuate according to a diffusion process. (Strong frequency dependence is required to prevent loss of variation at the selected locus in the absence of mutation: near the boundaries, $S \rightarrow p^{-\alpha}$ as $p \rightarrow 0$, $\alpha > 1$.) However, even this uncoupling into a set of single-variable equations does not lead to an explicit solution. The identities could be found by a change of timescale in the standard coalescent, but this does not lead to closed-form solutions (DONNELLY and KURTZ 1999b).

Weak random drift: If random drift is weak relative to mutation, recombination, and selection, then allele frequencies will be close to a deterministic limit. Many treatments of the effects of selection on genealogies assume this limit and apply the standard structured coalescent (*e.g.*, KAPLAN *et al.* 1988, 1989; HUDSON and KAPLAN 1995; WAKELEY 2001). For the model presented here, this limit corresponds to dropping the diffusion terms (*i.e.*, those terms involving ∂_p or $\partial_{p,p}$ from Equation 1) and thereby assuming that the population has always had the same allele frequency. This approximation is not explicitly dependent on the strength of selection, but does depend on both mutation and recombination rates, which determine the rate of mixing between allelic classes. Under this approximation, the mean pairwise coalescence time is

$$\bar{\tau} = \frac{pqR + p^2q^2(1 + R) + 4(Uc_2 + pqR)^2 + U(c_3 + 3p^2q^2)}{(Uc_2 + pqR) + 4U^2(c_1c_2 - \overline{pq}pq) + 8URc_2pq + 4p^2q^2R^2},$$

$$\text{where} \quad c_j = (\bar{p}q^j + \bar{q}p^j)\,. \tag{15}$$

A similar expression can be obtained for the identity in allelic state, which allows calculation of higher moments of the distribution of coalescence times. The overall mean, $E[\bar{\tau}]$, can be estimated either by integrating over the stationary distribution or by fixing p at its deterministic equilibrium value (NEUHAUSER and KRONE 1997).

Figure 10 shows the mean coalescence time as a function of the strength of balancing selection. Here, $U = N$, $\mu = 0.5$ or 0.25, and so effects are not large: balancing selection has a substantial effect on genealogies only when mutation is low enough that genes rarely move between backgrounds. The mean coalescence time does approach the prediction from Equation 15 for large S (right side of figure), but balancing selection must be extremely strong for the deterministic limit to be accurate. That is, weak random fluctuations in allele frequency can substantially reduce coalescence times. Note that weak disruptive selection (*e.g.*, underdominance) reduces coalescence times slightly, because allele frequencies tend to sweep back and forth between alternative alleles (see left of graph). However, strong underdominance has little effect, because the population is then near fixation.

Figure 11 shows an example in which fluctuations significantly reduce mean coalescence time despite strong selection. With $S = Ns = 32$, allele frequencies cluster around the equilibrium of $p_0 = 0.7$ (bell-shaped curve). The coalescence time is almost independent of allele frequency, simply because populations away from equilibrium are recently derived from populations close to equilibrium (thin solid curve). In contrast, the deterministic prediction (dashed line) ignores the diffusion of populations between different allele frequencies and so depends more strongly on allele frequency. Taking the value of this approximation at $p_0 = 0.7$ gives a substantial overestimate of mean coalescence time, even under such strong selection.

Figures 12 and 13 show similar results, but for mutation/selection balance rather than for balancing selection. The lines to the right in Figure 12 show the pre-

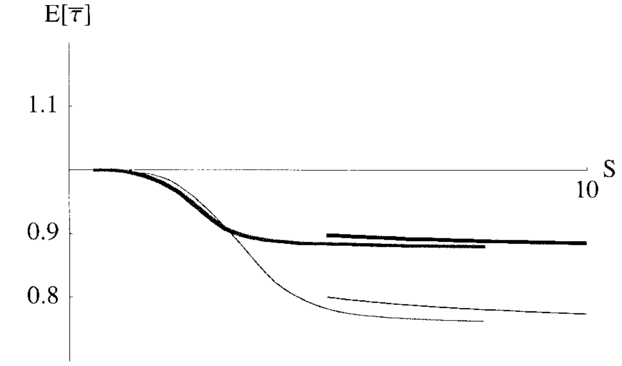

FIGURE 12.—The effect of purifying selection on the mean coalescence time, $E[\bar{\tau}]$. The two curves are for equilibrium frequency $U/S = 1/8$ (thick), $1/4$ (thin). There is no recombination ($R = 0$). The lines on the right are the predictions assuming that allele frequency is fixed at U/S.

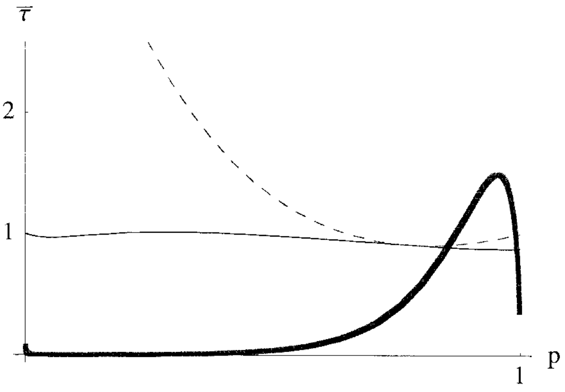

FIGURE 13.—The mean coalescence time, $\bar{\tau}$, as a function of allele frequency, at mutation-selection balance, $U = 3/8$, $\bar{p} = 0.001$, $R = 0$, $S = 3$. The thin solid curve gives the exact solution from Equations 5. This is compared with the prediction from Equation 15, which assumes that allele frequency is fixed; this prediction is independent of selection. The thick curve shows the stationary distribution.

diction for the deterministic limit (Equation 15); as $S \to \infty$, mean coalescence time is reduced by a factor $(1 - (U/S) + O(1/S^2))$ (CHARLESWORTH *et al.* 1993). This effect of "background selection" against deleterious alleles is reduced substantially by random drift for $S = Ns < {\sim}3$. Figure 13 shows the stationary distribution at $S = 3$, $U/S = \mu/s = 1/8$; as for balancing selection, a moderate degree of fluctuation around the deterministic expectation substantially reduces the effect of selection in reducing mean coalescence time [here, from $1 - (U/S) = 0.875$ at $S \to \infty$ to 0.908 at $S = 3$].

Rapid mixing: If mutation and/or recombination are strong relative to drift (U or $R \gg 1$), then there will be little divergence between allelic classes: Δ and θ will be small, and $\bar{f}$ close to $1/(1 + 4V)$. First, consider the case with U, $S \sim 1$ and $R \gg 1$. From Equations 8, we see that θ is augmented by the term $(1 - \bar{f})/pq$ and dissipated by recombination, and so is $O(1/R)$. Δ is generated by θ and dissipated by recombination and also mutation at the boundaries ($pq \sim 1/R$). Hence, $\Delta \sim 1/R^2$. The perturbation to $\bar{f}$ is therefore only of order $1/R^2$. Letting $\bar{f} = 1/(1 + 4V) + \phi$ the leading terms are

$$\begin{aligned}
0 &= -(1 + 4V)\phi + \left(2Spq + 2(\bar{p} - p)U + \frac{pq}{2}\partial_p\right)(\partial_p\phi - 4\Delta) + 4pqS\Delta \\
0 &= -2\left(U\frac{(q - p)}{pq}(\bar{p} - p) + R\right)\Delta + ((p - q)\theta - 2Spq\theta - 4(\bar{p} - p)U\theta - pq\partial_p\theta) \\
0 &= -4R\theta + \frac{(1 - \bar{f})}{pq}\cdot
\end{aligned} \tag{16}$$

Hence

$$\theta = \frac{V}{Rpq(1 + 4V)}$$

$$\Delta = -\frac{V(Spq + 2(\bar{p} - p)U)}{(U(\bar{p} - p)(q - p) + Rpq)R(1 + 4V)}\cdot \tag{17}$$

Note that the term $U(\bar{p} - p)(q - p)$ in the denominator of the expression for Δ is negligible except near the boundaries, when $Rpq \sim 1$. The perturbation ϕ to average identity is not given explicitly. However, its expected value, integrated over the stationary distribution, is given by Equation 10. Hence

$$\begin{aligned}
E[\bar{f}] &= \frac{1}{1 + 4V} + \frac{4}{1 + 4V}E[Spq\Delta[p]] \\
&= \frac{1}{1 + 4V} - \frac{4V}{(1 + 4V)^2}E\left[\frac{Spq(Spq + 2(\bar{p} - p)U)}{(U(\bar{p} - p)(q - p) + Rpq)R}\right].
\end{aligned} \tag{18}$$

Because the stationary distribution is known explicitly, this perturbation can readily be calculated. Moreover, it has a simple form when seen as a function of V. This allows us to take the inverse Laplace transform, which shows that the distribution of coalescence times is

$$\begin{aligned}
E[\Phi[\tau]] &= \text{Exp}[-\tau]\left(1 + (\tau - 1)E\left[\frac{Spq(Spq + 2(\bar{p} - p)U)}{(U(\bar{p} - p)(q - p) + Rpq)R}\right]\right) \\
E[\bar{\tau}] &= 1 + E\left[\frac{Spq(Spq + 2(\bar{p} - p)U)}{(U(\bar{p} - p)(q - p) + Rpq)R}\right].
\end{aligned} \tag{19}$$

Note that the key term $E[\ ]$ in Equation 19 is negative for purifying selection and positive for balancing selection.

If balancing selection is strong enough that the stationary density near the boundaries is negligible, and if $R \gg U$, then the term involving mutation in the denominator is negligible, and

$$\begin{aligned}
E[\Phi[\tau]] &\sim \text{Exp}[-\tau]\left(1 + (\tau - 1)E\left[\frac{S^2pq}{R^2}\right]\right) \\
E[\bar{\tau}] &\sim 1 + E\left[\frac{S^2pq}{R^2}\right].
\end{aligned} \tag{20}$$

Thus, when linkage is loose the effect of selection is to increase mean coalescence times by an amount proportional to the additive variance in fitness ($2S^2pq$). This can be understood as an inflation in the rate of random genetic drift due to inherited variation in fitness (HILL and ROBERTSON 1966; SANTIAGO and CABALLERO 1995). With balancing selection, the marginal selection

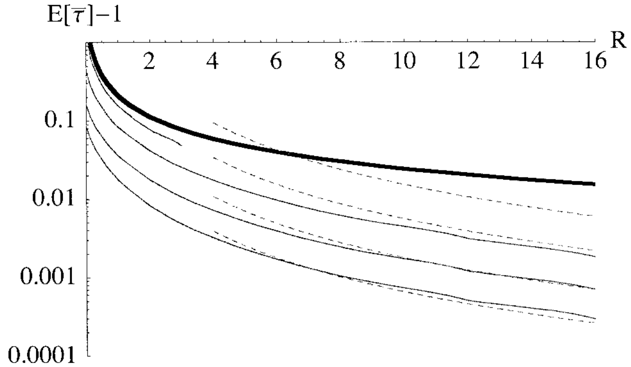

FIGURE 14.—The effect of balancing selection on mean coalescence time, plotted against recombination rate, R; $U = 0.05$. The vertical axis shows increases over the neutral value, $E[\overline{\tau}] - 1$, on a logarithmic scale. The top thick curve shows the deterministic limit, in which allele frequency is assumed to be fixed at $p_0 = 0.7$ (Equation 15). The thin solid curves are for $S_b = 4, 8, 16, 32$ (bottom to top), calculated using Equations 1 and 11. The dashed curves show the high recombination limit (Equation 19).

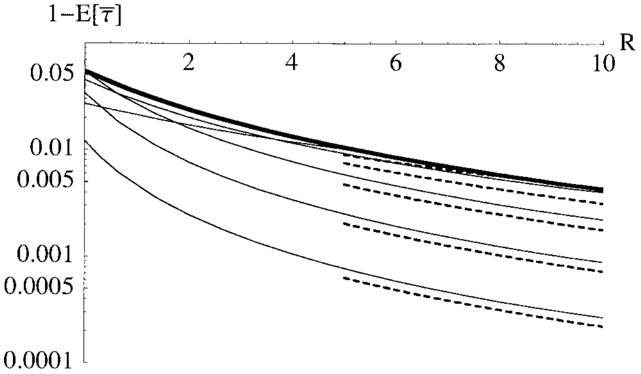

FIGURE 15.—The effect of purifying selection on mean coalescence time, plotted against recombination rate, R; $U = 0.5, \overline{p} = 0.5$. The vertical axis shows decreases from the neutral value, $1 - E[\overline{\tau}]$, on a logarithmic scale. The top thick curve shows the deterministic limit for $S = 8$, in which allele frequency is assumed to be fixed at $p_0 = 1 - (U/S)$ (Equation 15). The thin solid curves are for $S = 0.5, 1, 2, 4, 8$ (right side, bottom to top), calculated using Equations 1 and 11. The dashed curves show the high recombination limit (Equation 19).

coefficient $S = S_b(p_0 - p)$ is zero at $p = p_0$. However, allele frequency fluctuates around this expectation with variance var$(p) \sim 1/4S_b$ for large S_b. Hence, $E[S^2pq/R^2] \sim S_bp_0q_0/4R^2$. Note that this is not the same as the limit of Equation 15 as $R \rightarrow \infty$, which is $1/4R$. Allele frequency fluctuations have a significant effect that cannot be neglected even for strong selection.

We examine the accuracy of these approximations by considering the effect of increasing recombination. Figure 14 shows the increase in mean coalescence time caused by balancing selection, plotted against recombination rate. Mutation is set to a small positive value ($U = 0.05$), to ensure that a stationary distribution exists. However, mutation has a negligible effect on the results unless linkage is tight ($R \sim U$). The increase over the neutral value, $E[\overline{\tau}] - 1$, is plotted on a log scale, because when R is large, small effects must be discerned. As selection increases ($S_b = 4, 8, 16, 32$, bottom to top), mean coalescence time converges to the deterministic limit (thick line; Equation 15) However, convergence is slow for large R. There, the large R approximation of Equation 19 is accurate (dashed lines to right). However, the approximation is good only for $R > 10$, in which case mean coalescence time is increased by at most 2.5%. The large R approximation is not helpful for parameters that give a large effect.

Figure 15 shows a similar plot for the decrease in mean coalescence time caused by purifying selection. Now, mutation is set to a moderately high level ($U = 0.5$); with weak mutation, the population would almost always be fixed, and effects on linked variation would be negligible. The deterministic limit of Equation 15 (top thick line) now performs poorly. This is because it is based on the assumption that allele frequency is close to the deterministic equilibrium of $1 - (U/S)$, which is never the case for $U = 0.5$, even when selection is strong. This approximation is expected to be accurate only for $U \gg 1$. The large R approximation of Equation 19 is more accurate (dashed lines), but systematically underestimates the effect by ~18%. Examination of the differences in mean coalescence time between backgrounds, Δ^{τ}, θ^{τ}, shows that the approximation of Equations 17 breaks down near the margins ($q \sim (1/R)$). Since the stationary density is appreciable in this region for $R = 10$, much larger recombination rates would be needed for accuracy to be improved. As for balancing selection, this approximation is accurate only in cases where the effect on coalescence time is small.

For tight linkage, the effect of purifying selection at first increases with selection, but then decreases (see Figure 15, left side). This can be seen more clearly in Figure 16, which shows the mean coalescence time as a function of $S = Ns$, with complete linkage. As selection becomes very strong, the rare allele is driven out of the population, and so the genealogy returns toward its form under the neutral coalescent.

Large genealogies: We have concentrated on numerical examples for pairwise coalescence time partly because computations are then much faster, but also because the effects of fluctuations in allele frequency, and hence of selection, are primarily on the deeper parts of the genealogy, when there are just a few ancestral lineages. Here, we use Equation 5 to consider the effect of purifying selection on the expected total length of a larger genealogy. Figure 17 shows how the expected length depends on allele frequency and on the composition of the sample. If five genes of type P are sampled, then the genealogy is much shorter when that allele is rare (line rising steeply from left to right); similarly, if

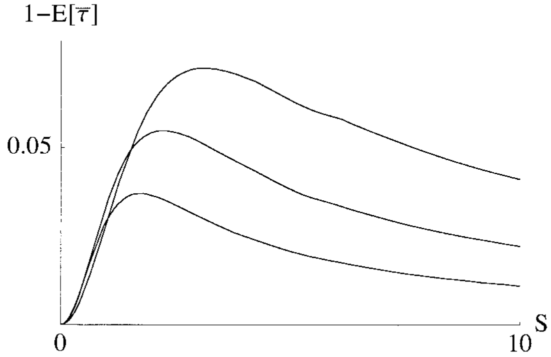

Figure 16.—The effect of purifying selection on mean coalescence time, plotted against selection, S, for $U = 0.25$, 0.5, 1; $\bar{p} = 0.5$. There is complete linkage ($R = 0$).

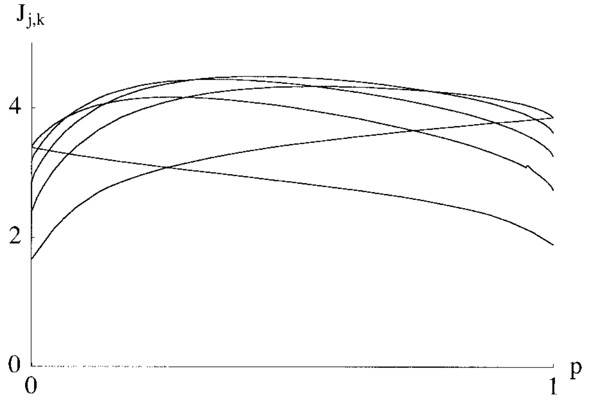

Figure 17.—The expected total length of the genealogy that connects a sample of five genes, plotted as a function of allele frequency; the six cases $J_{0,5}$ to $J_{5,0}$ are shown; $J_{0,5}$ increases most steeply from left to right, $J_{5,0}$ shows the opposite gradient, and the other four variables interpolate between these. There is purifying selection against Q alleles of strength $S = 2$; mutation is at rate $U = 0.5$ with $\bar{p} = 0.5$, and there is no recombination.

five copies of type Q are sampled, the genealogy is shorter when Q is rare, because coalescence is much more rapid within the rarer class. However, the relationship with allele frequency is quite weak for mixed samples, because coalescence occurs more slowly within each allelic class, and so lineages are likely to move between classes by mutation. In this example, there is purifying selection $S = Ns = 2$. However, the patterns for a neutral locus, or for other kinds of selection, are similar.

Figure 18 shows the net effect of purifying selection, $S = Ns$, on the expected total length, for up to $n = 50$ genes. Mutation rate is $U = N\mu = 0.5$, and there is no recombination. Equation 14 shows that the net effect of selection can be separated exactly into a sum, the jth term being due to the time when j lineages are present. Thus, Figure 18 shows the total effect on genealogies with 50 genes (top points) and the component effects on genealogies with up to 2, 3, 4, 5, . . . genes (bottom series of points). As explained above, most of the perturbation due to selection accrues when just a few lineages are present. Overall, the effects are small relative to the expected total tree length under neutrality (8.96 for $n = 50$ genes). The curves on the right give the simple approximation for large S, that effective population size and hence tree length is reduced by a factor $(1 - (U/S))$. This approximation is accurate only for S so large that the net effect is small.

DISCUSSION

We have used the equations set out by Kaplan *et al.* (1988; see also Hudson and Kaplan, 1988) to find the effects of weak selection ($Ns \sim 1$) on genealogies at a linked neutral locus. These genealogies are produced by the structured coalescent process, in which genes move between the genetic backgrounds defined by the selected locus as a result of both recombination and mutation of the selected alleles. Most previous studies have assumed that allele frequencies evolve deterministically and so are restricted to strong selection ($Ns \gg 1$). Here, we allow allele frequencies at the selected locus to evolve as a diffusion process. This leads to a set of coupled equations, each being the sum of two terms: a diffusion that allows for random changes in allele frequency through time and a recursion that describes the structured coalescent conditional on allele frequency. We find that quite small stochastic fluctuations in the frequencies of alternative genetic backgrounds substantially reduce the effects of selection.

In this article, we have considered only selection on a single biallelic locus and have for the most part sampled two genes from the linked neutral locus. Moreover, we have assumed that the selected locus has reached a stationary state, so that properties of the genealogy such as mean coalescence time can be taken to be functions of allele frequency only. In principle, it is straightforward to extend the method. Nonstationary processes can be described by taking the variables to be functions of time as well as allele frequency and following them using the same backward diffusion as in Equations 1. For example, one might ask about the genealogy immediately after a substitution has occurred by chance: Is the genealogy shortened in the same way as with a selective sweep?

Allowing multiple loci or more alleles under selection requires that the variables be functions of genotype frequencies, and this would greatly slow numerical calculations. More variables would also need to be followed, because genes can find themselves in many more genetic backgrounds. Extension to larger samples of genes at the neutral locus is simpler (Kaplan *et al.* 1988). Now, we follow the relationship between a set of j genes in background Q and k genes in background P; for a sample of n genes, this requires $(n - 1)(n + 4)/2$ variables. This does not raise substantial difficulties, be-

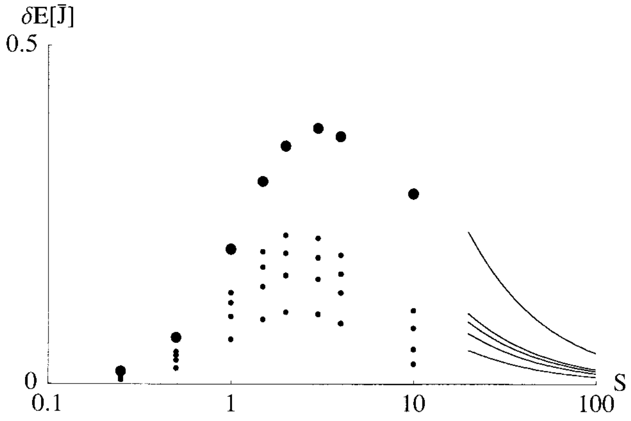

FIGURE 18.—The effect of purifying selection, *S*, on the expected total length of a genealogy, $\delta E[\bar{J}]$. The graph shows the reduction in $E[\bar{J}]$ below the neutral value, $\Sigma_{j=1}^{n-1}(1/j)$. The top large dots are for $n = 50$ genes; the bottom series of smaller dots are for $n = 2, 3, 4, 5$ genes. For comparison, $E[\bar{J}] = 8.96$ for $n = 50$ genes and 2, 3, 3.67, 4.17 . . . for 2, 3, 4, 5 . . . genes. The curves on the right are based on a reduction in effective population size by $(1 - U/S)$; this prediction applies when *S* is large. Mutation rate is $U = 0.5$, with $\bar{p} = 0.5$.

cause one can break up the calculation into sets of equations involving 2, 3, . . . *n* genes at a time. Direct solutions of the differential equations (Equation 5) are feasible up to five or so genes, and much larger numbers can be approximated by assuming that coalescence from *n* genes down to approximately five genes occurs quickly, relative to the timescale of the diffusion process (Figure 9).

Most existing results assume that allele frequencies evolve deterministically (*e.g.*, KAPLAN *et al.* 1988, 1989; HUDSON and KAPLAN 1995; WAKELEY 2001). Our numerical results show that this deterministic approximation is accurate only for quite strong selection: moderate fluctuations, reflected in dispersion of the stationary distribution, are sufficient to reduce effects substantially below those predicted. An obvious extension for the future is to make an expansion in powers of $1/Ns$, so as to improve on the deterministic approximation. The opposite approach is to examine the effects of weak selection. KRONE and NEUHAUSER (1997; Theorem 4.26) show that the effect of purifying selection and symmetric mutation on coalescence times is $O((Ns)^2)$ (see Figure 12). This result was obtained by showing that various terms, each corresponding to an alternative topology of the ancestral graph, cancel. Under our approach, we see immediately from Equation 19 that the first-order contribution is zero: to leading order, Δ is an odd function centered on $P = 0.5$, and so $E[Spq\Delta] = 0$ for a symmetrical stationary distribution. (It is not possible for us to compare with Krone and Neuhauser's Theorem 4.19, because this gives the probability that two individuals are identical in the sense that they have experienced no mutations that alter their allelic class since they shared a common ancestor. This is not a property of the genealogy alone and is not the same as the classical "identity by descent").

We see the main advantage of our method as being amenable to analytical approximations that might allow progress in understanding more complex cases. In principle, it would be possible to couple a diffusion process to the structured coalescent so as to generate the distribution of genealogies conditional on observed data (for example, observations of neutral mutations carried by sampled genes). However, this would be computationally impractical for more than a single selected locus and so would restrict attention to very simple hypotheses. Our hope is primarily to find general results that will help us to understand how diverse evolutionary processes influence sampled sequences.

We have shown that the expected change in certain properties of a genealogy caused by selection is equal to the covariance between those properties and fitness. Applying ROBERTSON's (1966) "secondary theorem" [or PRICE's (1970) equation] in this way is unusual in several respects. First, the argument applies to the expected value of a randomly distributed variable. Second, the argument is applied to samples of genes, rather than to individuals. Third, the phenotype of the sample of genes is taken to be a measure of their joint ancestry—in this instance, the expected length of the genealogy. However, one can see that PRICE's (1970) arguments do apply with these extensions: a sample of genes can be connected with their offspring in the next generation, and the fitness of the whole sample can be seen as covarying with its genealogical properties. By including the transmission terms in PRICE's (1970) equation, one can derive the whole of Equation 14: the expected state of the offspring sample changes as a result of the extra time step (first term) and as a result of coalescence (second term). It may be fruitful to apply this stochastic extension of Price's equations to other problems.

Recent simulations suggest that the effects of selection at one site on the genealogy are surprisingly weak (GOLDING 1997; NEUHAUSER and KRONE 1997; PRZEWORSKI *et al.* 1999; WILLIAMSON and ORIVE 2002). Of course, the frequencies of selected alleles are strongly affected: purifying selection eliminates allelic variation, and balancing selection maintains it. Thus, observations on biased codon usage can give direct estimates of *Ns* (MCVEAN and CHARLESWORTH 2000), because the frequencies of alternative bases at the third position are themselves under selection through their effects on translation. However, the structure of the genealogy is typically affected very little by selection, and thus, observations of this genealogy, or of closely linked neutral variation, tell us little about the action of selection. This is a serious practical limitation, because we often do not know the actual nucleotides that cause fitness differences and so must base our inferences on synonymous or noncoding variation that we assume to be neutral.

Przeworski *et al.* (1999) describe the "competition of alleles on a genealogy" whose structure is largely independent of the distribution of the selected alleles. This description is somewhat misleading, because the genealogy does depend strongly on the allelic state of the sample, even when no selection is acting (*e.g.*, Figure 1): genes in the same allelic state are more likely to share a recent common ancestor. However, when genealogies are averaged over the possible allelic configurations and over the stationary distribution of allele frequencies, the result is close to the neutral coalescent. Since we typically do not know which alleles influence fitness, we can usually observe only this average.

To examine the claim that selection has small effects on the genealogy, even at the selected locus, we must distinguish balancing from purifying selection. In the former case, there can be very strong effects provided that mutation rates are extremely low ($U \ll 1$), since this allows time for the two genetic backgrounds to diverge considerably. Such divergence can of course be observed directly, for example, at self-incompatibility loci in plants or the major histocompatibility locus in mammals (Hughes 1999). Extension of this argument to multiple balanced polymorphisms shows that in a sufficiently large population, and with sufficiently stable selection, coalescence times can become extremely long. However, random drift in even large populations greatly reduces this effect (Barton and Navarro 2002), because balancing selection cannot maintain every *combination* of selected alleles at constant frequency. Similarly, our single-locus results (Figure 10) show that even with $Ns_b = 20$, the effect on neutral variability can be more than halved. In reality, fluctuations in selection are likely to further reduce the effect of balancing selection below the ideal case of an extremely large population under constant conditions.

Recent discussions have concentrated on the effects of selection against deleterious mutations on genealogical structure. Purifying selection is of most general importance, because it acts on all functional sequences and because its effects can be substantial in aggregate, at least for organisms with a high genomic mutation rate and in regions of low recombination (Charlesworth *et al.* 1993). Purifying selection on a single site has surprisingly weak effects. For example, Williamson and Orive (2002, Table 4) found that with a total mutation rate $U = 5$, the expected total length of a genealogy connecting 50 genes is reduced by at most 28%, at $S = Ns \sim 5$. (We express U in terms of our model of reversible mutation; Williamson and Orive used $4N\mu_{Q\to P} = 4N\mu_{P\to Q} = 4U\bar{p} = 10$; their $2N\sigma$ is equivalent to our $2S$.) Przeworski *et al.* (1999) used a much lower mutation rate ($U = 0.1$) and observed a reduction of at most 0.53% at $S = 3$. However, because of the computational difficulties of simulations of either the whole population or the ancestral selection graph, the view that purifying selection at a single site has small effects on genealogies is based on quite limited numerical results.

We can identify three distinct reasons why purifying selection should have little effect on genealogies. First, even when selection is strong relative to drift, the effect of a single site is small. With complete linkage and two selected alleles, effective population size is reduced by a factor $(1 - (\mu/s))$ for large Ns (Charlesworth *et al.* 1993). This is a consequence of the fact that if back mutation is negligible, and if the fittest class is to be maintained indefinitely, then only the fraction $(1 - (\mu/s))$ of the population that is free of deleterious mutations can contribute in the long term. Since deleterious alleles are likely to be rare at any one site, the effect on neutral variability will be small. Nevertheless, the cumulative effects of many sites can be large. Averaging over a genetic map of length R, the effective population size is reduced by $\mathrm{Exp}[-\Sigma_i\, \mu_i/R]$, a factor that depends on the total mutation rate per map distance (Hudson and Kaplan 1995).

Second, for a given mutation rate, a significant effect is seen only for intermediate selection strengths. For strong selection, deleterious mutations become negligibly rare, while for weak selection, the effect is only of second order in S (Figure 12; Neuhauser and Krone 1997). The maximum possible effect increases with mutation rate (Figure 16), because the frequency of deleterious mutations is increased. However, this effect is offset by the more frequent movement of genes between the two alternative backgrounds, which reduces the covariance between the allele frequency in the sample, and the structure of the genealogy. For fixed and large S, mean coalescence time decreases in proportion to U (Figure 16, right side), but for weak selection, it is almost independent of mutation rate (Figure 16, left side). Presumably, the effect of increased frequency of deleterious alleles and of faster mixing between backgrounds counterbalances when S is small.

Finally, the effect of selection on large genealogies is appreciable only deep in the tree, when just a few lineages segregate. This is because lineages rapidly coalesce down to a small number of ancestors, and so the fluctuations in allele frequency that mediate the effects of selection have little influence during this period. Consider the perturbation to the expected total length of the genealogy, which is plotted against the strength of purifying selection in Figure 18, for $U = 0.5$. For 50 genes, the greatest reduction below the neutral value is by 4.2%, when $S = Ns \sim 3$. About half of this reduction is due to the effects of selection during the time when fewer than six lineages are present. Overall, the proportionate reduction in tree length is about the same for small and large samples. (In this example, pairwise coalescence time is reduced by at most 5.5%.) Statistics such as the length of external branches are expected to be much less sensitive to the effects of selection, an argument supported by the simulations of Przeworski *et al.* (1999) and Williamson and Orive (2002, Tables 3 and 4).

The methods that we have developed in this article suggest several avenues for future research. First, to what extent does selection distort the topology of the genealogy, rather than just changing its length? An understanding of such distortions is necessary if we are to be able to distinguish the effects of different kinds of selection from their effects on neutral variability. Second, analytical approximations can be developed for strong selection and for large genealogies, which may allow a better understanding of the joint effects of multiple loci on genetic variability. Finally, it may be possible to develop the idea that selection can act on samples of genes, and their genealogical relationships, in the same way that it acts on individual genes and their phenotypes. This suggests an intriguing link between the separate literatures on the stochastic evolution of genealogical relationships and on the deterministic evolution of groups of related individuals.

We are grateful to the Erwin Schrödinger International Institute for Mathematical Physics in Vienna for support during winter 2002/2003, where this manuscript was completed. We also acknowledge support from the Engineering and Physical Sciences Research Council and Biotechnology and Biological Sciences Research Council (grant MMI09726) and thank the referees for their helpful comments.

LITERATURE CITED

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. **72:** 123–133.

Barton, N. H., and A. Navarro, 2002 Extending the coalescent to multilocus systems: the case of balancing selection. Genet. Res. **79:** 129–139.

Barton, N. H., A. M. Etheridge and A. K. Sturm, 2003 The distribution of coalescence times in a fluctuating population. Ann. Appl. Probab. (in press).

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Darden, T., N. L. Kaplan and R. R. Hudson, 1989 A numerical method for calculating moments of coalescent times in finite populations with selection. J. Math. Biol. **27:** 355–368.

Donnelly, P. J., and T. G. Kurtz, 1999a Genealogical processes for Fleming-Viot models with selection and recombination. Ann. Appl. Probab. **9:** 1091–1148.

Donnelly, P. J., and T. G. Kurtz, 1999b Particle representations for measure-valued population models. Ann. Probab. **27:** 166–205.

Donnelly, P. J., M. Nordborg and P. Joyce, 2001 Likelihoods and simulation methods for a class of non-neutral population genetics models. Genetics **159:** 853–867.

Fearnhead, P., 2001 Perfect simulation from population genetic models with selection. Theor. Popul. Biol. **59:** 263–281.

Golding, G. B., 1997 The effect of purifying selection on genealogies, pp. 271–285 in *Progress in Population Genetics and Human Evolution* (IMA Volumes in Mathematics and Its Applications, Vol. 87), edited by P. Donnelly and S. Tavare. Springer-Verlag, New York.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

Hudson, R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Hudson, R. B., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

Hughes, A. L., 1999 *Adaptive Evolution of Genes and Genomes*. Oxford University Press, Oxford.

Kaplan, N. L., T. Darden and R. B. Hudson, 1988 The coalescent process in models with selection. Genetics **120:** 819–829.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The hitchhiking effect revisited. Genetics **123:** 887–899.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Krone, S. M., and C. Neuhauser, 1997 Ancestral processes with selection. Theor. Popul. Biol. **51:** 210–237.

McVean, G. A. T., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected sites on patterns of molecular evolution and variation. Genetics **155:** 929–944.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

Nagylaki, T., 1989 Gustave Malécot and the transition from classical to modern population genetics. Genetics **122:** 253–268.

Neuhauser, C., 1999 The ancestral graph and gene genealogy under frequency-dependent selection. Theor. Popul. Biol. **56:** 203–214.

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics **145:** 519–534.

Price, G. R., 1970 Selection and covariance. Nature **227:** 520–521.

Przeworski, M., B. Charlesworth and J. D. Wall, 1999 Genealogies and weak purifying selection. Mol. Biol. Evol. **16:** 246–252.

Robertson, A., 1966 A mathematical model of the culling process in dairy cattle. Anim. Prod. **8:** 95–108.

Rouhani, S., and N. H. Barton, 1987 Speciation and the "shifting balance" in a continuous population. Theor. Popul. Biol. **31:** 465–492.

Santiago, E., and A. Caballero, 1995 Effective size of populations under selection. Genetics **139:** 1013–1030.

Schulman, L., 1981 *Techniques and Applications of Path Integration*. John Wiley & Sons, New York.

Slade, P. F., 2000a Simulation of selected genealogies. Theor. Popul. Biol. **57:** 35–50.

Slade, P. F., 2000b Most recent common ancestor probability distributions in gene genealogies under selection. Theor. Popul. Biol. **58:** 291–305.

Slade, P. F., 2001 Simulation of 'hitch-hiking' genealogies. J. Math. Biol. **42:** 41–70.

Slatkin, M., 1996 Gene genealogies within mutant allelic classes. Genetics **142:** 579–587.

Slatkin, M., 2001 Simulating genealogies of selected alleles in a population of variable size. Genet. Res. **78:** 49–58.

Wakeley, J., 2001 The coalescent in an island model of population subdivision with variation among demes. Theor. Popul. Biol. **59:** 133–144.

Williamson, S. M., and M. E. Orive, 2002 The genealogy of a sequence subject to purifying selection at multiple sites. Mol. Biol. Evol. **19:** 1376–1384.

Wolfram, S., 1996 *The Mathematica Book*. Wolfram Media/Cambridge University Press, Cambridge, UK.

Communicating editor: W. Stephan

APPENDIX: NUMERICAL METHODS

Darden *et al.* (1989) described a numerical method for solving Equation 1 and gave some results for the case of two genes. Because they did not specify boundary conditions, they could not use standard algorithms. Instead, they approximated the differential equations on a discrete grid and thus obtained a set of linear equations that could be solved using matrix methods. This method is similar to solving the exact Wright-Fisher model for a finite population of $2N$ genes (see Figure 1).

Since we have specified the boundary conditions (Equations 2), we can solve the equilibrium version of

Equation 1 using the built-in algorithms in *Mathematica* (Wolfram 1996), at least for up to five or so genes. We proceed iteratively. Initially, all identities are set to zero. At the nth stage of the iteration we solve the stationary version of Equation 1 subject to the boundary conditions of Equations 2, using a "shooting method"; the boundary conditions and the contribution from mutation and recombination involve identities that are provided by our trial solution. Thus, for example, to solve for $f^n_{0,2}$, we set the boundary condition $f^n_{0,2}(0)$ to the value given by Equations 2 with $f^{n-1}_{1,1}$ and choose two different (negative) values of $\partial f_{0,2}(0)/\partial p$, α and β, say. This gives solutions $f^n_{0,2,1}$ and $f^n_{0,2,2}$. Write $z[\alpha]$ and $z[\beta]$ for the corresponding values of $(1 + 4V) f_{0,2}(1) + 2U\overline{q}\partial_p f_{0,2}\,(1)$. Since the correct choice of $\partial f_{0,2}(0)/\partial p$, γ, say, must give $z[\gamma] = 1$, and since we are dealing with a linear equation, the solution that we seek is given by setting $\gamma = ((1 - z[\beta])/(z[\alpha] - z[\beta]))f^n_{0,2,1} + ((z[\alpha] - 1)/(z[\alpha] - z[\beta]))f^n_{0,2,2}$. The iteration is repeated until the mean square change in the estimates is less than some small threshold.

For a sample of two genes, this procedure is successful after only a few iterations. For larger samples, the differential equations are less stable, and so it is necessary to integrate over a series of separate intervals. The algorithm we used is as follows. Starting at some small $\varepsilon \ll 1$, integrate forward as described above. By making a small change in the initial condition, obtain a second solution that is close to the first for small p, but that may diverge as p becomes large. Choose a point p_1 at which these two trial solutions are close to each other; we thus have an accurate solution for $\varepsilon < p < p_1$. Repeat the procedure starting from the opposite boundary ($p = 1 - \varepsilon$), obtaining a solution valid for $p_2 < p < 1 - \varepsilon$. If $p_2 < p_1$, splice the two solutions together at the point where they differ least. Otherwise, repeat the procedure, working in from p_1 to p_3, from p_2 to p_4, and so on until the solutions starting from left and right overlap.

With more than five or so genes, the differential equations become extremely sensitive to the values near the boundaries, and so "shooting methods" fail. This instability arises because the terms due to coalescence grow quadratically with the number of genes and so become extremely large relative to the diffusion terms that smooth the solution. This suggests an alternative approximation. We begin by calculating the identities among up to (say) five genes using the methods described above. To calculate identities among $n = 6$ genes, we first discard the diffusion terms (*i.e.*, the last two terms in Equation 1). This is equivalent to assuming that allele frequency fluctuations are negligible over the short timescale set by coalescence among the n lineages. This gives a set of linear equations: $0 = f_5 + B.f^0_6$, where B is a matrix, and f_5 is the vector of identities among sets of five genes, which we have already calculated. A correction f^1_6 to this solution f^0_6 can now be calculated by including the diffusion terms $0 = \mathrm{B}.f^1_6 + \mathcal{L}\partial_p f^0_6$, where $\mathcal{L} = 2(U(\overline{p} - p) + Spq) + (pq/2)\partial_p$. By making repeated corrections in this way, we obtain an asymptotic series $f^0_6, f^1_6, f^2_6. \ldots$ In numerical calculations, the first few terms converge toward the correct solution (calculated as above), but further corrections lead to divergence. A satisfactory approximation can be found by taking just the first correction, f^1_n. We have checked that this is close to the solution calculated using the method described above, at least for small n. Once a solution for n genes is found, the procedure can be repeated for $n + 1$ genes, and so on. The method is fast, and allows calculations for up to 100 or so genes. One further modification is required. Slight inaccuracies in solutions for a few ($n \sim 5$) genes tend to accumulate as fine-scaled fluctuations in solutions for more genes. We therefore force smooth solutions by fitting an eighth-order polynomial solution for each n.